Title:

Statistical Analysis of the Tables in Mahadevan's Concordance of the Indus Valley Script

Sole Author:

Michael Philip Oakes, RIILP, University of Wolverhampton, Wolverhampton, England.

Correspondence:

Michael Philip Oakes, MC137, RIILP, University of Wolverhampton, Stafford Street, Wolverhampton WV1 1LY, England. Michael.Oakes@wlv.ac.uk Tel:+44 (0)1902 322967 Fax: +44 (01902)

Keywords:

Mahadevan's Concordance; Indus Valley script; LNRE models; Pearson's Residuals; Correspondence Analysis.

Abstract

The Indus Script originates from the culture known as the Indus Valley Civilization which flourished from approximately 2600 to 1900 BC. Several thousand objects bearing these signs have been found over a wide area of Northern India and Pakistan. In 1977 Iravatham Mahadevan published a concordance of all of the scripts that had been discovered so far. Accompanying the concordance are a set of 9 tables showing the distribution of individual signs by position, archaeological site, object type, field symbol (accompanying image), and direction of writing. Analysis of the frequencies of the signs found so far using Large Numbers of Rare Events (LNRE) models enabled the total vocabulary of the language, including signs not yet found, to be about 857. All the tables were analysed using Pearson's residuals, and it was found that the signs were not randomly distributed, but some showed statistically significant associations with position, object, field symbol or direction of writing. A more detailed analysis of the relation between signs and field symbols was made using correspondence analysis, which showed that certain signs were associated with the unicorn symbol, while others were associated with the gharial and dotted circle symbols.

Keywords

Mahadevan, Concordance, Indus Valley script, LNRE models, Pearson's Residuals, Correspondence Analysis.

1. Introduction

Locklear (2017) writes that the Indus Script originates from the culture known as the Indus Valley Civilization, which flourished from 2600 to 1900 BC. The first seal bearing these signs was found by General Alexander Cunningham in 1872. Since then, several thousand of

these artefacts have been found over a wide area of Northern India and Pakistan. Many of them consist of a single line of signs above an illustration or "field symbol", typically a beautifully-drawn animal, such as in the example shown in Figure 1.

{INSERT FIGURE 1 ABOUT HERE}

About 70 of the symbols in the sign inventory resemble matchstick men carrying objects, while a further 40 or so are simple collections of strokes as if used for counting. Others resemble simple pictographs of animals, most commonly fish or birds. No real progress has been made towards their actual decipherment, but statistical tests involving information theoretic measures have shown that they are not merely random sets of symbols. However, it is not known for certain whether they constitute writing, in the sense that they encode the sounds or words of a human language. Some researchers, starting with Farmer, Sproat and Witzel (2004) think they may be simply religious or political symbols, like totem poles or coats of arms, or family names. The patterns occur mainly on seals , but have also been found on sealings, miniature tablets, pottery, copper tablets, bronze implements and ivory or bone rods (Mahadevan, 1977: 30). Reasons that the signs have so far proved impossible to decipher include a) the fact that the inscriptions are extremely short, on average 5 signs long; b) we do not know the underlying language; c) the Indus culture no longer exists and d) we have no Rosetta stone-like bilingual texts.

Most scholars tend to think the signs are linguistic, but Rao (reported by Locklear, 2017) believes that until we find longer sequences of signs or a multilingual text, statistical strategies are our best hope for understanding more about the Indus signs. Yadav et al.

(2017) found that inscribed objects found in different regions have noticeably different sign sequences, such as when comparing finds from Iraq with those from India/Pakistan. They also found some association between sign sequences and the medium of writing. Locklear (2017) also quotes Gabriel Recchia as saying that "there are significant differences between artefacts that appear in different sublocations within a site". Mahadevan's (1977) Concordance contains raw data for a number of potentially interesting associations between the sign frequencies and other factors, including archaeological site and the type of inscribed object. We will examine each of these using Pearson's residuals in Section 4.

2. Is the Indus Script writing?

Rao et al. (2009) used the measure of conditional entropy to compare the Indus sign sequences with other human languages and non-linguistic symbols such as DNA and Fortran computer code, and found that in this respect they more closely resembled the human languages. Sproat's (2014) counter argument was that conditional entropy just tells you that as sequence is not completely random, and this is known already. Sproat has created a corpus of both non-linguistic and linguistic symbols, where the 7 non-linguistic sign systems include totem poles, sequences of five-day forecast icons from a weather corpus website, the Kudurrus Mesopotamian deity symbols, and a subset of Indus Valley "bar seal" texts. The 14 linguistic systems include Amharic, Arabic, Chinese, English, Hindi and Tamil. Chinese showed the lowest entropy of all the systems, the Kudurrus (unlike its behaviour in the simulation of the Kudurrus stones by Rao et al.) showed the highest, with all the other systems (including the Indus bar seals) somewhere in the middle, with linguistic systems overlapping with non-linguistic. Thus conditional entropy did not discriminate well between linguistic and non-linguistic systems.

Sproat found that Lee et al.'s (2010) C_r measure and repetition rate (both of which are related to sentence length) were better discriminators between linguistic and non-linguistic systems than conditional entropy. In fact, a Wilcoxon signed rank test showed that out of a number of linguistic features as candidates for distinguishing linguistic vs non-linguistic symbols only C_r and repetition rate were statistically significant.

The formula for C_r is as follows:

$$C_r = \frac{N_d}{N_u} + a\frac{S_d}{T_d}$$

where N_d and N_u are the number of bigram types and the number of unigram types respectively, a is a constant (found empirically by Lee et al. to be 7) S_d is the number of bigrams that occur once (hapax legomena) and T_D is the number of bigram tokens. Lee found that if $C_r >= 4.89$, the system was linguistic.

Boy Scout merit badges (another example of a non-linguistic system with some language-like properties) are never earned twice – therefore repetition rate could be a discriminator between linguistic and non-linguistic symbols. The Indus signs have has low repetition within a single inscription, but where repetition does occur, the same sign can appear up to four times in a row. (Sproat, 2014:467-469). As an example of repetition rate, consider the sequence A A B A C D B (Sproat, 2014:468). Here the repetition rate = r / R = 1 / 3. R is 3 because there are 2 As after the first one plus one B after the first one, r is 1 because there is one 2-gram of As, and there are no other immediate adjacencies. Sproat (2014: 469) found that repetition rate was the "cleanest separator" of his corpus data into linguistic versus non-linguistic. Repetition rate was negatively correlated with mean text length (Pearson's r = -0.49), so the shorter the text, the less chance of repetition. Although C_r was only correlated with text length with r = 0.39, when the outlier Amharic is removed from the data set r increased to 0.71.

In his definition of writing, Sproat does not include meaning-bearing systems that do not directly encode language (2014: 474). He feels (Sproat, 2014: 478) that the mean length of utterance should be considered as the most basic feature in distinguishing linguistic and non-linguistic systems. Since the Indus sign sequences are so short, it is unlikely to be a true linguistic script.

The first problem that will be tackled in this paper is that of estimating the total vocabulary of the Indus signs, using the data in Mahadevan's concordance, specifically the table of "Frequency and positional distribution of signs" (Mahadevan, 1977: 717-723). Here 417 sign types have been identified so far, and the frequency of each sign in the corpus of signs discovered up to 1977 is given. We will use this data to estimate the number of signs there are altogether in the language of the Indus signs, both in the corpus and still to be found. We will first review the history of techniques which have sought to extrapolate from the frequencies in a sample discovered so far to estimate the total population size. All of these techniques could potentially be used to estimate the total size of the Indus sign vocabulary, but we chose to use the family of Large Numbers of Random Events (LNER) models, since these have been implemented on Baroni and Evert's (2014) ZipfR package.

3. Estimation of vocabulary size

In this section we describe ecological studies which have sought to extrapolate data of the numbers of species of animals found to date, to estimate the number of species "out there" which remain to be discovered. These techniques have been used to estimate the total vocabulary (in words, or other linguistic features) of an author, or language, based on the frequencies of that feature in an existing finite corpus.

Paxton (1998) collected descriptions of discoveries of large sea animals (>2m in length) from the scientific literature for the period 1758 to 1995, and found 117 new species were

described in that time period. He assumed that description and discovery rates were sufficiently correlated for the former to be a reliable proxy of the latter. The cumulative discovery curve (number of species discovered so far against time) was approximated by a hyperbola of the form

$$S_n = S_{max}\left(\frac{BS_n}{n}\right)$$

where n is the year since observations began, S_n is the number of new species described by year n, S_{max} is the total number of species existing, both discovered and yet to be discovered; and B is a constant found by fitting the theoretical curve to the empirical data, and is the approximate mean number of years between successive discoveries of new species. B gave the best fit between the empirical discovery data and the theoretical hyperbola when B was about 5.3. Substituting these values into the formula suggested that about 47 species remained to be discovered.

Solow and Smith (2005) fitted a non-stationary Poisson process to Paxton's data from 1929. A Bernouilli trials process is the discrete counterpart of the Poisson process. If a dice is thrown at discrete time intervals, we can record the number of throws between each successive six. In a Poisson process, events can occur at any time, not just at set intervals. Examples are the length of time between accidents at a traffic location, or the length of time between replacements of a light bulb. In a stationary Poisson process, the average rate at which events occur remains constant, but in a non-stationary Poisson process the average rate varies with time, such as in the case of passengers arriving to use public transport, which is much greater during rush hours than at other times. The rate of discovery of new species of large sea animal is a non-stationary Poisson process. As is the case with encountering new word types in a text, the rate of discovery of new species is highest at the start of the recording process, but falls once most of the common species have already been found.

Another factor, noted by Paxton (1998), is that with the decline of whaling and the number of sailing ships, there are now fewer opportunities to sight new species.

If a graph is plotted of the "cumulative description record", or the number of species record by each date in the past, then by fitting a theoretical "ideal" curve over the observation period, we can extrapolate this curve to estimate how many new species will be found by a given date in the future. If this curve is asymptotic, and tends towards a maximum value of a certain number of species described, we can guess how many different species there are altogether, both discovered and yet to be discovered.

The model has two parameters, θ, where 1/θ is the mean rate of discovery of new species at the beginning of the observation period, and β is related to the rate of decay in the rate of discovery of new species as follows: g(t) is the effort and skill required to sight a new species at time t, and is equal to exp(β t). β and θ are found by maximising the product over all species of the probability density function (PDF) at time j and the cumulative density function (CDF) at time 0. Since the first observation was in 1829 and the last in 1995, t_0, the end of the observation period, was taken to be 169 years. T_j is used to denote the time of the first sighting of species j. Estimates of theta and beta were 52.6 and 0.013. Now we can work out CDF of T_j:

$$prob(T_j \leq t_j) = 1 - \frac{\theta}{\theta + \frac{1}{\beta}(\exp(\beta\, t_j) - 1)} = P(t_j)$$

The estimate of m, the number of unknown species, is given by

$$\hat{m} = n\,\frac{1 - \hat{P}\,(t_0)}{\hat{P}(t_0)}$$

where $\hat{P}$ (t_0) is the estimate of $P(t_0)$ found by first estimating θ and β. The estimated probability that a previously undiscovered species is first found during the observation period

was $\hat{P}$ (t_0) = 0.92, and the estimate $\hat{m}$ of undiscovered species at the end of the observation period was 10.2. This suggests a total population of large oceangoing animals of m + n = 10.2 + 117 = 127.2.

This model is an extension of the Jelinski-Moranda model (Jelinski and Moranda, 1972) of the rates at which bugs in a computer program are uncovered. As time goes on, discovery of new errors requires more effort so we need the function g(t) to account for this trend.

In the approach by Efron and Thisted (1976), instead of estimating the numbers of species of animals, we estimate the vocabulary of Shakespeare. Individual word tokens correspond to individual animals, word types correspond to species. The original observation period is represented by the canon of Shakespeare works, time is represented by the number of words since the start of the canon, and t is the length of the canon, which is 884,647 words. Using the method of Good and Toulmin (1956), they counted how many word tokens were used just once in the canon (14,376), twice (4343), three times (2292), and so on.

Assuming that new word types were encountered in a Poisson process, Good and Toulmin found that it is possible to estimate the number of new words that would occur in an additional text of length t, $\hat{\Delta}$ (t), as follows:

$$\hat{\Delta}(t) = n_1 t - n_2 t^2 + n_3 t^3 \ldots$$

where n_x is the number of words encountered exactly x times in the canon. Thus if a new body of Shakespeare's work were discovered, of the same length as the canon, t would be 1 and $\hat{\Delta}$ (t) would be 11,430. If $t \geq 1$ the formula does not converge, but Good and Toulmin use Euler's transformation to enforce convergence of the series. To find the total extent of Shakespeare's vocabulary, we need to estimate $\hat{\Delta}$ (t) when t is infinity. This gave Efron and Thisted an estimate of 35000 words in addition to those already in the canon.

Another scenario related to the questions of how many members of a species are there, and what is the total vocabulary of an author, is the question of how many sides there are on a traditional die. Each hitherto unseen face of the die is like a new species or a new word, so we can talk about the vocabulary growth curve (VCG) which is the number of different faces seen so far, against the number of throws of the die. Baayen, (2001:52) shows the mean VCG for up to 100 throws of a fair die. Figure 2 in this paper shows similar data for a simulation in which the number of faces seen so far for each number of throws N in the range 0 to 100 was determined 1000 times and the mean values E[V(N)] plotted with black dots. After a small number of throws, the curve is very close to (but does not quite reach) its asymptote of 6. The curves below made with white dots show the various spectral frequencies, which are E[V(1,N)], the proportion of times we have found exactly one face so far, E[V(2,N)], the proportion of times we have seen exactly two different faces so far, and so on. These spectral frequency curves all have the characteristic that they reach a maximum value, and then fall back towards zero. Most random events produce vocabulary curves with these two characteristics. In contrast, word frequency distributions are characterised by Large Numbers of Rare Events (LNRE), an idea developed by Khmaladze (1987) and expanded upon by Baayen (2001: 51-57). Baroni and Evert (2014:7) give the following example of an LNRE event: the number of different Italian words with the prefix "ri-" (roughly corresponding to "re-" in English) found in a corpus. The results of their study are shown in Figure 3. In the upper, bold line, we see the V(N), the number of word types (different examples of words starting in "ri-" encountered so far) as a function of N, the number of word tokens (all words, not just those starting in "ri-") from the start of the corpus. In the lower curve, drawn in lighter type, V(1,N), the number of words beginning in "ri" which have been seen exactly once so far is plotted against N. The characteristics of these curves differ from those of the corresponding curves for the non-LNRE events shown in Figure 2. Firstly the top line in

Figure 3 never reaches anywhere near its asymptote, even though the corpus is very large. Secondly, the spectral element V(1,N) never reaches a maximum value. These observations are typical of LNRE events. Word frequency distributions are LNRE distributions, because there are many individual words each with an individually low frequency of occurrence. For example, in the British National Corpus (BNC), more than half of the word types have a relative frequency of .0000001 (Baayen, 2001:55).

{INSERT FIGURES 2 AND 3 HERE}

As was the case with the estimation of the number of species of large oceangoing animals, we can extrapolate beyond the end of the observation period (the length of the corpus) to estimate the full extent of the vocabulary of a language. For this, we can use a number of LNRE models, such as the three parametric models implemented in the ZipfR toolkit by Baroni and Evert (2014). These models are the Generalised Inverse Gauss Poisson (GIGP) (Baayen, 2001:89-93), Zipf-Mandelbrot (ZM) (Evert, 2004) and the Finite Zipf-Mandelbrot (FZM) (Evert, 2004). All three models assume independent Poisson sampling (Evert, 2004:2). The GIGP has three free parameters, γ, b and c. In the ZM model, the vocabulary growth curve follows the following relation, known as Heaps' law or Herdan's law: $E[V(N)] = C' \cdot N^{\alpha}$, where C' and α are free parameters determined empirically. The ZM model assumes an infinite vocabulary, which is unrealistic for natural language data. Thus the FZM model was developed, where only word types with a probability more than a threshold A are included in the vocabulary.

In this paper we use each of these three models to extrapolate the frequencies of each character type listed in Mahadevan's (1977) concordance of the Indus Valley texts, to estimate the number of different signs (discovered and undiscovered) there are in that language. Previous work (Oakes, 2016) produced this estimate for the Indus script, starting

with the appendix of 677 Indus signs with their frequencies in Bryan Wells' "Epigraphic Approaches to Indus Writing" (2011). A major difficulty with working with lost languages such as the Indus script in general is that there is often no universal agreement on the number of symbols discovered so far. Robinson (2009:284) states that

"Computerized analysis is a good idea in principle, but it is potentially misleading if based on a doubtful sign list. We certainly cannot rely on a computer to make judgements about which signs are allographs (variants of the same sign) and which are ligatures (combinations of two or more simple signs)."

In this paper we use Mahadevan's concordance as a starting point, which identifies the smaller number of 417 symbols in the sign inventory discovered so far. The input data to each of the LNRE models in the ZipfR package by Baroni and Evert (2014) is in the form of a frequency spectrum, such as the one shown in Table 1. Vm is the number of characters seen in the corpus exactly m times, so in the concordance 112 symbols are seen exactly once, 47 are seen twice, and one symbol is seen 1395 times. The frequency spectrum was derived from Table I of Mahadevan's concordance, entitled "Frequency and Positional Distribution of Signs", lists for each of the discovered signs in the Indus script the number of occurrences in the corpus where it occurs in an inscription as a "solus" (on its own), in the initial position, in a medial position, in the final position, and the total number of occurrences. The values in this final column were collated using a small Perl program to produce the frequency spectrum shown in Table 1.

{INSERT TABLE 1 ABOUT HERE }

The frequency spectrum was read in by the ZipfR package by using the R command:

```
m.spc = read.spc(file=file.choose())
```

and then the three LNRE models were run in turn as follows:

```
m.gigp = lnre("gigp", m.spc, exact=FALSE)

m.zm = lnre("zm", m.spc, exact=FALSE)

m.fzm = lnre("fzm", m.spc, exact=FALSE)
```

The results for each of the three models (viewed by typing `m.gigp`, `m.zm` and `m.fzm` respectively) are shown in Table 2.

{INSERT TABLE 2 ABOUT HERE}

Table 2 first shows the results for the GIGP LNRE model. The parameters of the model, γ, B and C, which give the best fit between the emprical data and the theoretical curve are shown first. Baayen (2001:123) states that the downhill simplex method of Nelder and Mead (1965) is especially useful for parameter optimisation. Zipf size (Baayen, 2001:80) is a characteristic constant of a corpus, and it is the sample size N at which the harmonic distribution holds best. This distribution relates the frequency of words occurring exactly m times, V(m,N), with the total vocabulary size V(N) at that point, as is given by

$$E[V(m,N)] = \frac{V(N)}{m(m+1)}$$

The next piece of data in the output is the estimated population size predicted by the GIGP model, which corresponds to the estimated total number of characters in the Indus sign inventory, taking into account the 417 symbols found so far, and extrapolating to estimate the number of symbols which have not yet been discovered. The parameters estimated from the size of the corpus (13372 characters) result in a table of observed and expected values. The observed values are the actual number of character types actually counted in the corpus (V),

the number of characters appearing exactly once (V1) and so on. The expected values are those predicted by the GIGP model with the parameters set at the above values. To evaluate the goodness-of-fit of the actual to the theoretical data, a standard chi-squared test would take the observed and expected (V(m,N) and E[V(m,n)] spectral values for the first r elements, and the observed and expected frequencies of all subsequent elements combined. The resulting chi-squared value would be the sum of the quantity $(O - E)^2 / E$ over all r + 1 pairs of values. (Baayen: 2001:118). However, we should also take into account the fact that the various spectrum elements have substantially significant variances, as described by Baayen (2001:119-122). The resulting total chi-squared value relates to statistical significance. If the corresponding p value is more than the arbitrary threshold of 0.05, there is no significant difference between the actual data and the curve predicted by the LNRE model.

The three models give widely differing estimates for the entire size of the Indus sign inventory. GIGP predicts about 857 symbols, ZM an infinite number, and FZM about 578 symbols. The p values for the goodness-of-fit for each model to the empirical data are considerably less than 0.05 for each model, showing that all three models differ significantly from the actual data. However, the largest p value for all three LNRE models, 0.00065, was obtained from the GIGP model, showing that this model fitted the empirical data better than the other two (FZM gave a p-value of about 1.28×10^{-9}, and ZM gave a p-value of about 1.23×10^{-23}) and hence GIGP can be taken to give the most reliable estimate of the size of the complete Indus character set, both already discovered and still to be discovered, of about 857 symbols). Figure 4 shows the vocabulary growth curve (VGC) predicted by the GIGP model. The R commands to produce this plot were as follows, where the lines show the size of the existing corpus and the number of different symbols discovered so far:

```
m.gigp.vgc = lnre.vgc(m.gigp,(1:100) * (0.1 * N(m.spc))
```

```
plot(m.gigp.vgc)

lines(x=c(0,150000), y = c(417,417))

lines(x=c(13372,13372), y=c(0, 800))
```

{INSERT FIG. 4 ABOUT HERE}

4. Pearson's Residuals for tabular data

A second analysis was performed on the data in Table I of Mahadevan's concordance, to find whether there was a statistically significant association between any of the Indus signs and their position (initial, medial or final) in the corpus. The chi-squared test was used, considering only those signs which had expected frequencies of at least 5 for all three positions. The remaining signs were grouped into a single "other" category. Altogether there are 13,182 sign tokens in the corpus which occur in sign groups of more than one character. Of these, 3010 had been found in the initial position, 7196 in a medial position, and 2976 in the final position. The number of initial symbols is different to the number of final symbols because some symbols are obscured and not included in this analysis. There were also 190 symbols which had been found as singletons, but these were not sufficient to produce any expected values over 5, and so were not included in the analysis. Using the relation (expected values = row total x column total / grand total), it was possible to select those signs which would have expected values greater than 5 for all three positions (initial, medial and final) as those where the row totals (sum of frequencies in all three positions) were greater than (13182 / 2976) x 5 ≈ 22. Using the R command `chisq.test` an overall chi-squared value was found for the entire data set, showing a highly significant association between symbol type and symbol position (X^2 = 9655.3, df = 190, $p < 2.2 \times 10^{-16}$). To find which of the symbols and their positions had contributed most to this overall chi-squared value, the

Pearson residuals were found for each cell in the original frequency table. The square of the Pearson residual is the contribution of each cell to Pearson's chi-squared statistic, so each residual is equal to $(O - E) / \sqrt{E}$ where O is the observed frequency and E is the expected frequency. Pearson's residuals relate to statistical significance in an equivalent way to z-scores. Thus a residual of 3.29 has a p-value of less than 0.001. The results of the chi-squared test had been placed into a data set called `m2` with the command `m2=chisq.test`; the Pearson residuals for each cell could then be found by typing `m2$residuals`. The most statistically significant residuals are tabulated in Table 1 below, and the corresponding raw frequencies are shown in parentheses.

{INSERT TABLE 3 ABOUT HERE}

From Table 1 we see that symbols 95, 267, 391 have the most significant association with the start position, and symbols 12, 15, 176, 211, 328 and 342 have the most significant association with the final position. Previously, Yadav et al. (2010) had shown that sign 267 is the most frequent in the start position, and sign 342 is the most frequent at the end. The unequal text beginner and ender distributions they found provided internal evidence for syntax.

Since by examining a whole table of Pearson residuals, we are effectively conducting a whole family of statistical tests on the same original contingency table. This increases our chance of making a Type 1 error, where the null hypothesis is incorrectly rejected. To compensate for this, we apply the Bonferroni correction. If we are looking to see which residuals are significant at $p < 0.001$, to apply the Bonferroni correction we divide this value by the

number of cells in the contingency table to obtain $p' = 0.001 / (96 * 3)$. To find the z-score corresponding to this value we used the R command qnorm(c(0.001 / (96 * 3))) which is about 4.50. According to this, all the symbols in Table 3 are significantly associated with one or more positions in an individual inscription. Table II in Mahadevan's concordance gives the raw frequencies of pairwise combinations of symbols, and is not analysed further in this paper.

Table III in Mahadevan's concordance, "Distribution of sites by signs", gives the number of times each symbol has been found at the two larger archaelogical sites, Mohenjodaro and Harappa and at five smaller sites which we will consider collectively. A total of 7821 characters have been found at Mohenjodaro, 4359 at Harappa, and 1192 at the minor archaeological sites. 73 different symbols gave expected values of more than 5 for each site (or for the smaller sites combined), and the remaining symbols were pooled into an "other" category. This produced an overall chi-squared value of 1624.4, df = 146, $p < 2.2 \times 10^{-16}$, showing that the distribution of symbols across sites was not random, but some symbols were found relatively more often at a particular site. Table 4 below shows the Pearson residuals for each symbol at each site for each of the cases where $p < 0.001$. Using the Bonferroni correction, this required a residual of at least 4.44. The Pearson residual values are followed by the raw frequencies of each symbol at each site. No symbol was significantly characteristic of the Mohenjodaro site, symbols 89, 95, 169, 176 and 328 were significantly associated with findings at the Harappa site, and symbols 1 and 99 were significantly associated with the smaller archaeological sites.

{INSERT TABLE 4 ABOUT HERE}

Table IV in Mahadevan's concordance, "Distribution of signs by object types" lists the types of object which have been found bearing Indus symbols are (1) seals, (2) sealings, (3) miniature stone, terracotta or faience tablets, (4) pottery, (5) copper tablets, (6) bronze implements, (7) ivory or bone rods, and (8) miscellaneous inscribed objects. In order to ensure expected values of at least 5 for every cell, we group object types (3) to (8) into a single, "other" category. All signs not occurring 27 times or more in total were also pooled into a single category. Altogether there were 8312 symbols found on seals, 2582 on sealings, and 2478 on other objects. The overall chi-squared value was 2285.3 with 172 degrees of freedom, giving a p-value $< 2.2 \times 10^{-26}$. Thus the symbols are not randomly distributed across the types of object they are found on, but certain symbols are more likely to be found written upon particular kinds of object. When using the Bonferroni correction, in order to be significant at $p < 0.001$, the residuals had to be more than 4.47. The 7 signs which had statistically significant association with object type (and their raw frequencies in parentheses) are shown in Table 5 below. Symbol 99 is significantly associated with seals, 176 and 328 with sealings, and 48, 89, 95, 176, 244 and 328 with other object types.

{INSERT TABLE 5 ABOUT HERE}

Table VI of Mahadevan's concordance shows the "Distribution of direction of writing by sites". We used this small table in its entirety, except for pooling the last two archaeological sites "Other Sites" and "West Asia". For the direction of writing, "Others" includes single sign, top to bottom, symmetrical and "doubtful". The overall chi-squared value was 114.44 for df = 10, giving a p-value < 2.2e-16. Thus there was a statistically significant association between direction of writing and archaeological site. Using the Bonferroni correction, for the residuals to be significant at $p < 0.001$ required them to have a value of at least 3.86. Examination of the residuals showed that the two cells which contributed most to the overall

chi-squared value were those for the Left to Right direction at the two largest sites: this direction was proportionally seen much more often at Harappa than Mohenjodaro. Table 6 shows Pearson's residuals for the association between direction of writing and archaeological site, with the raw frequencies (observed values) in parentheses.

{INSERT TABLE 6 ABOUT HERE}

Table VII of Mahadevan's concordance is "Distribution of direction of writing by object types". This table was used in its entirety, except the bottom three rows (bronze implements, ivory and bone rods, and miscellaneous objects) were pooled into the "others" category. The table had an overall chi-squared value of 246.53, with df = 10, giving a p-value < 2.2e-16. Thus the direction of writing was not random across the object types, but some object types were more likely to be inscribed with writing in a particular direction. Using the Bonferroni correction, for the residuals to be significant at $p < 0.001$ required them to have a value of at least 3.86. Examination of the residuals showed that the Left to Right direction was proportionally more rare on seals, but more common on miniature tablets. Table 7 shows Pearson's residuals for the association between direction of writing and object type, with the raw frequencies (observed values) in parentheses.

{INSERT TABLE 7 ABOUT HERE}

Table VIII of Mahadevan's concordance is "Distribution of field symbols by sites". Many of the Indus inscriptions are decorated with beautiful miniature pictures of animals, trees and leaves, anthropomorphic forms or other motifs. The most common depictions are unicorns, generally facing a cult object (symbol 01), humped bull (03), short-horned bulls generally with head lowered over a trough (04), elephant, sometimes with a trough in front (07), uncertain animal (mostly bovine) in the field (35), gharial, sometimes with a fish in its jaw or

surrounded by fish (36), or one or more dotted circles (83). The other field depictions were pooled into a single class. Similarly, in the analysis described in this paper, all archaeological sites apart from the two major ones were pooled into a single group. The overall chi-squared value for this table was 259.9 for df = 14, giving a p-value < 2.2×10^{-16}. Thus the distribution of field symbols across archaeological sites was not random. To find which associations between field symbol and site had contributed most to the overall chi-squared value, the Pearson's residual for each cell was examined. With the Bonferroni correction, all residuals with a value of 3.93 or more were significant at $p < 0.001$. The most significant residuals corresponded to the significant association between the gharial and the dotted circle field symbols with the Harappa site. There was also a negative association between the "other" field symbols and the "other" sites. Table 8 shows Pearson's residuals for the association between field symbol and archaeological site, with the raw frequencies (observed values) in parentheses.

{INSERT TABLE 8 ABOUT HERE}

Mahadevan's concordance also contains the table "Distribution of field symbols by object types". In the experiment described in this paper, all the object types apart from seals were pooled into a single category, and all symbols other than the most frequent 13 were also pooled into a single category. Symbols considered separately in this experiment, but not in Table 8 above, were: Rhinoceros, generally with a trough in front (11), Goat-antelope with a short tail (13), Fabulous animal with the body of a ram, horns of a bull, trunk of an elephant, hindlegs of a tiger and an upraised serpent-like tail (25), Indian Kino tree (*Pterocarpus marsupium*), generally within a railing or on a platform (44), different geometrical patterns generally occupying the whole field on one side of the inscribed object (97) and different ornamental borders of geometrical patterns at either or both ends of a text or along the edges

(98). A chi-squared test was performed, giving a chi-squared value = 953.01, df = 13, p-value < 2.2e-16. Thus the symbols are not randomly distributed across objects, but some are significantly associated with seals, and some with other objects. Using the Bonferroni correction, Pearson's residuals were statistically significant if greater than 3.97. This means that the unicorn symbol occurs significantly more often on seals, while the gharial, kino tree and dotted circles occur significantly more often on other objects. Table 9 shows Pearson's residuals for the association between field symbol and object type, with the raw frequencies (observed values) in parentheses.

{INSERT TABLE 9 ABOUT HERE}

In the final experiment we used Table V in Mahadevan's concordance entitled "Distribution of signs with field symbols". A contingency table was prepared with the 10 most frequent signs in the entire corpus (not just those appearing in conjunction with field symbols), all other signs being pooled into an "other" category. Since the frequency of signs with the unicorn symbol was much greater than the frequency of signs with other symbols individually, all symbols apart from the unicorn were also pooled into a single class. Overall a chi-squared value of 269 was obtained, for df = 10, giving a p-value < 2.2e-16. Thus the signs are not distributed randomly across the accompanying field symbols, but some have an affinity with the unicorn symbol while others have an affinity with the other symbols. Using the Bonferroni correction, Pearson's residuals were statistically significant if greater than 3.91. Thus the association between sign 99 and the unicorn symbol was significant, as were the associations between signs 176 and 328 with the other symbols. Table 10 shows Pearson's residuals for the association between signs and field symbols, with the raw frequencies (observed values) in parentheses. Due to the need for expected values to be at least 5, it was only possible to run this chi-squared experiment with a small number of signs

and only two symbol classes. In the following section we will again use the data in Mahadevan's "Distribution of signs with field symbols", this time to perform correspondence analysis, which enables us to consider all the signs and all the symbols occurring at least once in the corpus.

{INSERT TABLE 10 ABOUT HERE}

5. Correspondence Analysis of signs and field symbols

The data in Table V of Mahadevan's concordance "Distribution of signs by field symbols" was transformed (using a Perl program called `make.pl`) into a matrix (file `ca.txt`) where each row corresponds to one of the signs in the Indus corpus, and each column to one of the field symbols. Thus the value in cell [1,4] contains the number of times sign 1 (stick man) was found with field symbol number 4 (short-horned bull). The program also produced a vector of row labels (`labels.txt`). Correspondence Analysis was run as described by Baayen (2008: 128-135). Assuming that the `languageR` package is already installed, the commands were:

```
indus = read.table(file=file.choose())
labels = read.table(file=file.choose())
indus = indus[,-81] or indus = indus + 1
indus.ca  = corres.fnc(indus)
plot(ind.ca,  rlabels  =  labels$V1,  rcex=1.0,  extreme=0.1,
ccol="black")
```

Since there are no signs associated with symbol 81, and correspondence analysis does not allow column (or row) totals of zero, it was necessary to remove the column corresponding to symbol 81 with the command `indus = indus[,-81]`. However the resulting plot was

very difficult to interpret, with all the signs and symbols crowded along the two main axes. To produce a clearer plot, Laplace smoothing was applied using the command `indus = indus + 1`. This produced the plot shown in Figure 5, and shows that the main factors are co-occurrrence with the unicorn symbol (V1) and factor 2 is co-occurrence with symbols V36 (gharial) and V83 (dotted circles). The signs most closely associated with symbol V1 were 59 (co-occurring 169 times altogether in the corpus), 99 (385 times), 267 (212 times) and 342 (550 times). The signs most closely associated with Factor 2 were 171 (co-occurring 14 times with V36 and 7 times with V83), 176 (17 and 23 times respectively) and 328 (3 and 28 times respectively). Altogether there are 5491 sign tokens which co-occur with symbol V1, 252 which co-occur with V36, and 234 which co-occur with V83.

{INSERT FIG 5 ABOUT HERE}

6. Conclusion

In this paper we have used LNRE models and Mahadevan's concordance of discovered signs in the Indus script to estimate the total number of sign types in the language as a whole, including those as yet undiscovered. The total number of sign types found so far estimated by Mahadevan is 417. In this study, the best-fitting model to the empirical data was GIGP, and this gave an estimate of the total vocabulary of 857 signs. Mahadevan estimated that allowing a margin for allographs and undiscovered signs, "the present best estimate is 425 plus or minus 25 signs" (Robinson, 2009: 281). This study gives a much higher estimate of the number of yet undiscovered signs.

Another leading scholar of the Indus script is Asko Parpola. His first three of a planned four volumes of photographs have greatly stimulated the study of the script. His sign inventory (Parpola, 1994) contains 386 signs with 12 more unnumbered signs, just slightly fewer than

in Mahadevan's sign list, and is taken by most scholars to be definitive. Other estimates of the number of signs in the Indus sign inventory do not agree so well with Mahadevan and Parpola. Rao suggests a number of only 62 (Robinson, 2009:284). At the other extreme, the Appendix to Wells (2011) lists 677 different signs, and a previous study (Oakes, 2016) used this with the GIGP model to estimate a total vocabulary of 1396 signs, including those so far still undiscovered.

This indicates a major difficulty when working with corpora of lost languages with large sign inventories – it is often difficult to distinguish sign variants from distinct signs, and to distinguish single signs from ligatures, the joining of two smaller signs. This results in widely differing estimates of vocabulary size in the texts discovered so far, and is a major impediment to decipherment. However, Yadav and Vahia (2011) do set out criteria for sign classification and decomposition.

The experiments with Pearson's residuals showed that the distribution of the signs was not random throughout the corpus, but there were significant associations between the frequency and the positional distribution of signs; signs and archaeological sites; signs and the object types on which they are inscribed; signs and the field symbols (accompanying pictures), direction of writing and archaeological sites; direction of writing and object types. Correspondence Analysis gave a more detailed picture of which signs showed the most affinity with which field symbols in the corpus, and showed that the most distinctive field symbols in this respect were the unicorn, gharial and dotted circles.

Although there is no consensus on the meaning of the script, the number of signs in the inventory suggests a logo-syllabic script, with a large number of signs representing concepts, and a much smaller set representing syllables. However, there are other systems with several hundred signs in their inventories, such as the Old Cretan "hieroglyphic seals", the Naxi

Dongba symbols from Yunnan Province in China, and the Rongorongo from Easter Island., but this fact does not automatically translate them into mature phonetic writing systems. Most scholars think the Indus script is writing (Parpola, 1994; Vidale, 2007; Rao et al., 2009; Yadav et al., 2017) and the balance of opinion favours a Dravidian rather than a Sanskrit-related underlying language (Robinson, 2016). Mahadevan himself believes that the Indus script represents an early form of Dravidian (Robinson, 2009: 276).

Sproat (2014: 457) classifies information systems which use symbols to convey meaning as either non-linguistic (like traffic signs, the information conveyed is not tied to any specific human language) or a true linguistic writing system, where a particular language is being encoded, and symbols refer to specific phonemes, syllables, or even words. This definition of a writing system contrasts with Powell's broader definition which implies that "any conventional meaning-bearing system is writing" (Sproat, 2014:475). An intermediate position is taken by Baines (2008: 348-349) who states that writing systems bring together two factors that distinguish human beings from other animals: an elaborate material culture; and language. In practice, most writing systems are not extensively developed and cover only some of the topics covered by speech. Many create modes of communication that are substantially different from spoken ones. However, they do not include domains such as the purely visual or numerical or mathematical. Taking the views of Powell and Baines, the Indus signs would constitute a form of writing. However, if they do encode meaning, it is not clear whether they encode a human language or at least, a subset of it, (and thus are linguistic) or are purely symbolic (non-linguistic, in Sproat's definition). Although I favour Sproat's definition, the statistics described in this paper require no assumptions about whether or not the Indus scripts are writing, and can be

used to estimate the size of the sign inventory in any language or non-linguistic symbol system.

There are some similarities between the Indus seals and the Ancient Egyptian bone and ivory tags found in the cemetery of Umm el-Qa'ab at Abydos, which have been dated at about 3200 BC. They resemble the Indus seals in that each one is about 2 – 3 cm. square, although unlike the Indus seals, each one has a small hole in one of the top corners, to enable it to be attached to a box or jar – apparently holding commodities of high value – or to bales of cloth by a piece of cord. Each Abydos tag is inscribed with between one and four signs, so they are of similar length to the Indus scripts. The surviving Umm el-Qa'ab material is insufficient for various published readings of the text (Dreyer et al, 1998; Breyer, 2002; Kahl, 2003), despite their plausibility, to be more than theoretical. These readings suggest that the texts are a mixture of phonetic and logographic signs. Their purpose was to record the contents, place of origin, quantity, length or ownership of items to be stored in the royal tomb (Wilkinson, 2007). This could also be the function of the similarly brief Indus seals, which may record a subset of a language for economic purposes rather than a full language (Robinson, 2016).

At the time of writing, computers do not possess the intuition and flair that humans have for decipherment of unknown human-made languages or ciphers. Although computers are superior in terms of load of work, speed of computation, endurance and consistency, it is unlikely that significant breakthroughs in the interpretation of the Indus scripts will be made by computers alone, unless there are significant developments in artificial intelligence. Given

that the Indus scripts found so far are extremely brief, we may also have to wait until inscriptions appear which are sufficiently long to allow for a plausible elucidation or decipherment.

References

Baayen R. H. (2001). *Word Frequency Distributions.* Kluwer Academic Publishers.

Baayen, R. H. (2008). Analysing Linguistic Data: A Practical Introduction to Statistics using R. Cambridge.

Baines, J. (2008). Writing and its Multiple Disappearances, In: The Disappearance of Writing Systems. Perspectives on Literacy and Communication, edited by John Baines, John Bennet and Stephen Houston, London: Equinox.

Baroni M. and Evert S. (2014). The zipfR package for lexical statistics: A tutorial introduction. 3 October 2014.

Breyer, F.A.K. (2002). Die Schriftzeugnisse des prädynastischen Königsgraben U-j in Umm el-Qaab: Versuch einer Neuinterpretation. *Journal of Egyptian Archaeology*, 88: 53-65.

Dreyer, G., Hartung, U. & Pumpenmeier, F. (1998). Umm el-Qaab I: das prädynastische Königsgrab U-j und seine frühen Schriftzeugnisse. *Archäologische Veröffentlichungen (AV) 86*. Mainz.

Farmer S., Sproat R. & Witzel M. (2004). The collapse of the Indus script thesis: The myth of a literate Harappan civilization. *The Electronic Journal of Vedic Studies* 11(2): 19-57.

Good, I. J. & Toulmin, G.H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. Biometrika 43, 45-63.

Jelinski, Z. and Moranda, P. B. (1972) Software Reliability Research. In *Statistical Computer Performance Evaluation* (ed. W. Freiburger). London: Academic: 465-484.

Kahl, J. (2003). Die frühen Schriftzeugnisse aus dem Grab U-j in Umm el-Qaab. *Chronique d'Egypte*, 78: 112-135.

Khmaladze E.V. (1987). The statistical analysis of large number of rare events. *Technical Report* MS-R8804, Dept. of Mathematical Statistics, CWI. Amsterdam: Center for Mathematics and Computer Science.

Lee, R, Jonathan, P & Ziman, P. (2010). Pictish symbols revealed as a written language through application of Shannon. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 466(2121):2545-60.

Locklear, M. (2017). Cipher War. After a century of failing to crack an ancient script, linguists turn to machines. January 25th, 2017. http://www.theverge.com/2017/1/25/14371450/indus-valley-civilization-ancient-seals-symbols-language-algorithms-ai Accessed 19.6.17

Mahadevan, I. (1977) *The Indus Script: Texts, Concordance and Tables*. Delhi: Archaeological Survey of India.

Nelder J. & Mead R. (1965). A simplex method for function minimization. *Computer Journal* 7:308-313.

Parpola, A. (1994) *Deciphering the Indus Script*, Cambridge University Press.

Paxton C. G. M. (1998). A cumulative species description curve for large open water marine animals. *J. Mar. Biol. Assoc.* 78, 1389-1391.

Rao R., Yadav N., Vahia M. N., Joglekar, H., Adikhari, R. & Mahadevan I. (2009). Entropic evidence for linguistic structure in the Indus script. *Science* 324(5931): 1165. Online (with supplementary information) at http://homes.cs.washington.edu/~rao/ScienceIndus.pdf

Robinson, A. (2009). *Lost Languages*. Thames and Hudson.

Robinson, A (2016). Cracking the Indus script. *Nature* 526 : 499 – 501.

Solow, A. R. & Smith, W. K. (2005). On estimating the number of species from the discovery record. Proceedings of the Royal Society B 272, 285-287.

Sproat, R. (2014). A statistical comparison of written language and nonlinguistic symbol systems. *Language* 90(2):457-481. 2014.

Vidale, M. (2007). The collapse melts down: A reply to Farmer. Sproat and Witzel. *East and West* 57(1-4): 333-366.

Wells B. K., (2011). *Epigraphic Approaches to Indus Writing*. Oxbow Books.

Wilkinson, T. (2007) Ancient Writing: Early Hieroglyphs from Abydos, Ancient Egypt. In Brian M. Fagan, editor, *Discovery! Unearthing the New Treasures of Archaeology*. Thames and Hudson : 238.

Yadav, N., Joglekar, H., Rao, R.P.N., Vahia, M.N., Mahadevan, I. & Adikhari, R. (2010) Statistical analysis of the Indus script using n-grams. *PLOS One* 5(3).

Yadav, N. & Vahia, M. N. (2011). Indus Script: A Study of its Sign Design. *SCRIPTA*: *International Journal of Writing Systems*, 3: 133-172.

Yadav, N., Salgaonkar. A & Vahia, M. (2017). Clustering Indus texts using k-means. International Journal of Computer Applications 162(1): 16-21

Table 1. Frequency spectrum for the Indus signs.

m	Vm
1	112
2	47
3	21
4	26
5	13
6	10
7	14
8	15
9	6
10	4
11	11
12	2
13	11
14	5
15	4
16	2
17	2
18	1
19	2
20	2
21	4
22	5
23	1

24	2
25	1
26	4
27	3
29	3
30	1
32	1
33	1
35	5
38	3
41	1
42	1
43	1
44	1
48	1
49	1
50	2
51	2
53	2
54	2
57	1
58	1
59	1
60	1
61	1
63	2

64	1
69	1
70	3
73	2
76	1
78	1
80	1
88	2
89	1
90	1
91	2
92	1
93	1
99	1
102	1
105	2
118	1
126	1
130	1
132	1
134	2
136	1
149	1
151	1
168	1
170	1

177 1

188 1

193 1

195 1

207 1

212 1

216 1

227 1

236 1

240 1

279 1

314 1

323 1

355 1

365 1

376 1

381 1

649 1

1395 1

Table 2. Results for the vocabulary estimations for the Indus script using LNRE models.

```
> m.gigp
Generalized Inverse Gauss-Poisson (GIGP) LNRE model.
Parameters:
   Shape:          gamma = -0.2182712
   Lower decay:        B = 0.0165727
   Upper decay:        C = 0.02810703
 [ Zipf size:          Z = 35.57829 ]
Population size: S = 857.0086
Sampling method: Poisson, approximations are allowed.

Parameters estimated from sample of size N = 13372:
                  V     V1    V2    V3   V4    V5
   Observed: 417.00 112.00 47.00 21.00 26.0 13.00 ...
   Expected: 411.72 111.18 49.07 29.54 20.6 15.57 ...

Goodness-of-fit (multivariate chi-squared test):
         X2 df           p
   28.98998  9 0.0006505671


> m.fzm
finite Zipf-Mandelbrot LNRE model.
```

```
Parameters:
   Shape:          alpha = 0.3956673
   Lower cutoff:       A = 1.817636e-05
   Upper cutoff:       B = 0.06518818
 [ Normalization:      C = 3.169647 ]
Population size: S = 578.0451
Sampling method: Poisson, approximations are allowed.

Parameters estimated from sample of size N = 13372:
               V     V1    V2    V3    V4    V5
   Observed: 417 112.00 47.00 21.00 26.00 13.00 ...
   Expected: 417 113.63 57.02 32.33 21.16 15.26 ...

Goodness-of-fit (multivariate chi-squared test):
         X2 df            p
   57.75991  8 1.279598e-09



> m.zm
Zipf-Mandelbrot LNRE model.
Parameters:
   Shape:          alpha = 0.3452679
   Upper cutoff:       B = 0.06459507
 [ Normalization:      C = 3.936083 ]
Population size: S = Inf
Sampling method: Poisson, approximations are allowed.
```

```
Parameters estimated from sample of size N = 13372:

                V     V1    V2 V3    V4    V5

   Observed: 417 112.00 47.00 21 26.00 13.00 ...

   Expected: 417 143.98 47.13 26 17.25 12.61 ...


Goodness-of-fit (multivariate chi-squared test):

        X2 df            p

   126.917  8 1.231198e-23
```

Table 3. Pearson's residuals for the association between symbol and position in the Indus script.

Symbol	Initial	Medial	Final
12	-4.01 (1)	-5.20 (9)	12.11 (69)
15	-5.11 (1)	-4.53 (30)	12.19 (92)
95	11.61 (59)	-5.06 (5)	-3.80 (0)
176	-8.99 (0)	11.17 (38)	26.41 (316)
211	-7.18 (0)	-7.33 (42)	18.62 (184)
267	22.95 (298)	-9.97 (62)	-7.57 (15)
328	-6.36 (19)	-11.07 (29)	23.61 (274)
342	-17.77 (1)	-12.33 (420)	37.05 (971)
391	13.77 (135)	-6.23 (41)	-4.15 (16)
Other	14.57 (544)	-8.11 (488)	-2.04 (256)

Table 4. Pearson's residuals for the association between symbol and archaeological site in the Indus script.

Symbol	Mohenjodaro	Harappa	Other sites
1	-0.35 (70)	-3.27 (22)	5.91 (42)
89	-3.91 (120)	7.12 (175)	-3.40 (19)
95	-5.06 (5)	8.14 (58)	-2.53 (1)
99	1.08 (374)	-5.11 (137)	5.87 (138)
169	-3.77 (134)	6.49 (179)	-2.57 (27)
176	-6.65 (101)	11.48 (239)	-4.57 (15)
328	-11.16 (28)	17.83 (288)	-5.39 (7)

Table 5. Pearson's residuals for the association between symbol and object type in the Indus script.

	Seals	Sealings	Other
48	-3.46 (67)	0.26 (34)	5.78 (67)
89	-4.82 (124)	1.83 (75)	6.61 (115)
95	-5.91 (2)	0.17 (13)	10.13 (49)
99	6.10 (515)	0.94 (115)	-9.71 (19)
176	-8.61 (89)	6.31 (121)	8.80 (145)
244	-7.73 (54)	-4.98 (13)	18.38 (177)
328	-12.64 (19)	9.81 (140)	12.39 (164)

Table 6. Pearson's residuals for the association between direction of writing and archaeological site.

	Right to Left	Left to Right	Others
Mohenjodaro	2.45 (1533)	-6.16 (48)	-2.05 (149)
Harappa	-2.19 (1076)	5.99 (148)	1.45 (158)
Chanhudaro	-0.19 (70)	0.14 (6)	0.42 (10)
Lothal	0.33 (169)	0.55 (15)	-1.37 (14)
Kalibangan	-1.21 (87)	0.78 (10)	2.84 (22)
Other Sites / West Asia	-1.34 (39)	2.14 (8)	2.09 (11)

Table 7. Pearson's residuals for the association between direction of writing and object types.

	Right to Left	Left to Right	Others
Seals	2.74 (1749)	-5.31 (69)	-3.56 (150)
Sealings	-0.80 (601)	1.70 (61)	0.92 (84)
Miniature tablets	-2.77 (362)	9.37 (87)	0.39 (54)
Pottery graffiti	-3.15 (74)	0.54 (10)	8.57 (44)
Copper tablets	0.36 (139)	-1.43 (6)	0.12 (17)
Other	-0.80 (49)	-1.12 (2)	3.19 (15)

Table 8. Pearson's residuals for the association between field symbol and archaeological site.

Symbol	Mohenjodaro	Harappa	Other Sites	Total field symbol frequency
01 Unicorn	0.96 (747)	-3.37 (239)	2.72 (173)	(1159)
03 Humped Bull	2.14 (46)	-2.11 (6)	-1.78 (2)	(54)
04 Short-horned Bull	1.03 (67)	-2.70 (11)	1.61 (17)	(95)
07 Elephant	0.48 (37)	-2.42 (5)	2.45 (13)	(55)
35 Uncertain animal	-0.25 (37)	0.28 (17)	0.17 (8)	(62)
36 Gharial	-3.53 (11)	6.61 (36)	-1.62 (2)	(49)
83 Dotted Circles	-5.22 (8)	9.61 (57)	-2.15 (2)	(67)
Other	0.34 (287)	2.24 (140)	-4.03 (25)	(452)

Table 9. Pearson's residuals for the association between field symbol and object type.

Field Symbol	Seals	Other objects	Total Frequency of Field Symbol
01 Unicorn	8.44 (1045)	-12.74 (114)	(1159)
03 Humped Bull	2.20 (51)	-3.32 (3)	(54)
04 Short-horned bull	1.97 (82)	-2.97 (13)	(95)
07 Elephant	-0.36 (36)	0.54 (19)	(55)
11 Rhinoceros	-2.13 (16)	3.22 (23)	(39)
13 Goat-antelope	-2.60 (12)	3.93 (24)	(36)
25 Fabulous animal	-1.04 (10)	1.58 (10)	(20)
35 Uncertain animal	1.51 (53)	-2.27 (9)	(62)
36 Gharial	-4.81 (6)	7.26 (43)	(49)
44 Kino Tree	-4.45 (2)	6.72 (32)	(34)
83 Dotted Circles	-6.53 (2)	9.86 (65)	(67)
97 Geometrical pattern	-4.25 (0)	6.42 (26)	(26)
98 Ornamental border	-4.17 (0)	6.29 (25)	(25)
Other symbols	-8.65 (70)	13.07 (202)	(272)

Table 10. Pearson's residuals for the association between signs and field symbols.

Sign	Frequency with unicorn symbol	Frequency with any other symbol
342	-1.90 (550)	2.37 (431)
99	4.29 (385)	-5.35 (124)
59	0.72 (169)	-0.90 (94)
267	2.69 (212)	-3.34 (78)
87	-0.26 (129)	0.32 (88)
176	-5.12 (55)	6.37 (123)
328	-6.52 (11)	8.12 (92)
89	-1.52 (75)	1.89 (72)
67	1.27 (135)	1.59 (64)
169	-1.52 (65)	1.89 (64)
Others	0.79 (3705)	-0.98 (2310)